Sidestepping the combinatorial explosion: Towards a processing model based on discriminative learning

R. Harald Baayen (baayen@ualberta.ca)

Peter Hendrix (hendrix@ualberta.ca)

Department of Linguistics, 4-55 Assiniboia Hall, University of Alberta, Edmonton T6G 2E5, CANADA

Abstract

Arnon and Snider (2010) documented frequency effects for compositional 4-grams independently of the frequencies of lower-order n-grams. They argue that comprehenders apparently store frequency information about multi-word units. We show that n-gram frequency effects can emerge in a parameter-free computational model driven by naive discriminative learning, trained on a sample of 300,000 4-word phrases from the British National Corpus. The discriminative learning model is a full decomposition model, associating orthographic input features straightforwardly with meanings. The model does not make use of separate representations for derived or inflected words, nor for compounds, nor for phrases. Nevertheless, frequency effects are correctly predicted for all these linguistic units. Naive discriminative learning provides the simplest and most economical explanation for frequency effects in language processing, obviating the need to posit counters in the head for, and the existence of, hundreds of millions of n-gram representations. **Keywords:** naive discriminative learning; Rescorla-Wagner equations; n-gram frequency effects; computational modeling

Introduction

In a recent study, Arnon and Snider (2010) reported frequency effects for four-word n-grams (see also, e.g., Tremblay & Baayen, 2010). They take this finding as evidence that multi-word phrases are units of representation, just as frequency effects for regular morphologically complex words have been taken to imply the presence of representations for such words in the mental lexicon (Baayen, Dijkstra, & Schreuder, 1997).

A recent computational model proposed by Baayen, Milin, Filipović Durdević, Hendrix, and Marelli (2010) generates simulated processing latencies for complex words that correlate well with the whole-word frequencies of those words. Crucially, this model does not make use of any representations at all for complex words. The model's architecture comprises only two representational layers, a layer of orthographic units (letters and letter bigrams) and a layer of elementary meanings. Each orthographic unit is connected with each meaning. The weights on these connections are estimated from corpus-derived conditional co-occurrence matrices using the equilibrium equations (Danks, 2003) of the Rescorla-Wagner equations (Wagner & Rescorla, 1972) for discriminative learning. Baayen et al. refer to their model as instantiating *naive* discriminative reading as the weights on the connections to a given meaning are estimated independently of all other meanings, as in naive Bayes classifiers. The simulated latencies predicted by the naive discriminative reader (henceforth NDR) reflect not only whole word frequency effects, but also morphological family size effects, inflectional entropy effects, constituent frequency effects, and paradigmatic entropy effects (Milin, Filipović Durđević, & Prado Martín, 2009).

The weights of the NDR were calculated on the basis of the co-occurrence frequencies extracted from 1,496,103 different three-word sequences extracted from the British National Corpus. This made it possible for the NDR too also correctly model syntactic relative entropy effects present in single-word lexical decision latencies. Phrasal frequency effects were also predicted, but not tested. This is the primary goal of the present paper, which takes the materials of Experiment 1 of Arnon and Snider (2010) as its point of departure. The first question to be addressed is whether naive discriminative learning correctly predicts the observed phrasal frequency effect, without making use of representations for n-grams. The second question addresses the complexity of naive discriminative learning compared to models assuming independent representations for n-grams.

Simulation

For each of the 47 different final words of the 56 four-word phrases of Experiment 1 of Arnon and Snider (2010), all occurrences were retrieved from the British National Corpus, together with the three preceding words in the sentence, when available. From the resulting data set, those four-grams were selected that consisted only of letters, including the apostrophe. Next, all words in these phrases were lemmatized using the CELEX lexical database, as follows. First, inflected words were traced back to their uninflected base form. Second, if this uninflected base form was morphologically complex, the CELEX parse was used to retrieve its component formatives. In this way, the phrase *a British provincial city* was associated with the set of meanings {A, BRITAIN, ISH, PROVINCE, IAL, CITY}, and the n-gram *abnormalities take on many* with the set of meanings {AB, NORM, AL, ITY, TAKE, ON, MANY}. Phrases with one or more words for which a CELEX parse was not available were discarded. This left us with 337,069 different phrase types, representing 562,905 phrase tokens. It is noteworthy that just 47 words — the 47 different phrase-final words of Experiment 1 of Arnon and Snider (2010) —

generate no less than 337,069 different phrases covering 7494 distinct meanings.

The 'lexicon' of 337,069 phrases was used to calculate the weights from letters and letter bigrams to the meanings associated with the constituents of the phrases. Given a phrase as input, the weights on the connections of the active letters and letter bigrams to a meaning were summed to obtain that meaning's activation. The activation of a phrase was modeled as the sum of the activations of its associated meanings. A phrasal decision latency was taken to be proportional to the log of the reciprocal of this summed activation.

In the analysis of the simulated latencies, we considered phrase pair (henceforth Pair) to be a fixed-effect factor, rather than a random-effect factor, as the phrase pairs do not constitute a random sample from a larger population of such pairs. To the contrary, the pairwise matching procedure used by Arnon and Snider (2010) resulted in a very specific, non-random set of phrase pairs. For their set of phrase-final words, repeating their matching procedure would result in a very similar, if not identical, set of phrases. In other words, the factor levels of Pair are repeatable. Therefore, Pair was entered into the model specification as a fixed-effect factor.

Following Arnon and Snider (2010), we fitted a regression model to the simulated latencies with as predictors the frequency of the four-word phrase, the frequency of the last word, and the frequency of the last two words. All frequencies were calculated from the lexicon used to estimate the model's connection weight matrix, and log-transformed. Stepwise backward model selection using AIC resulted in a model with only the n-gram frequency and Pair as predictors. The slope for n-gram frequency was estimated at -0.018 ($t(27) = -2.189$, $p = 0.0374$). The presence of a significant effect for n-gram frequency and the absence of significant effects for the frequency of the fourth word and for the frequency of the final bigram, exactly mirrors the pattern of result reported by Arnon and Snider (2010) for the empirical phrase decision latencies. Importantly, the model generating the simulated latencies is parameter-free, and driven completely by its corpus input.

Table 1: Model comparison for a simple main effects model (model 1), a model with a multiplicative interaction of n-gram frequency by fourth word frequency (model 2), and a model using a tensor product to model this interaction (model 3).

	res. df	res. dev	df	dev	F	p
model 1	25.00	0.06				
model 2	24.00	0.06	1.00	0.00	1.08	0.31
model 3	21.92	0.03	2.08	0.03	11.47	0.00

Thus far, each meaning associated with the phrase was given equal weight. Equal weights may not be optimal, however. For instance, Baayen et al. (2010) observed for compounds that the weight of the head meaning was best modeled as half that of the weight of the modifier meaning. For phrases in a phrasal decision task, equal weights may likewise not properly reflect the task demands. As more words become available to the participant, the next word becomes more predictable, and hence should have a decreased weight for a yes-response. We implemented this conceptualization of the phrasal decision task by proportionally decreasing the weight for each successive meaning. For an n-gram such as *you like to try*, the weight for the first word was set to 1, for the second word it was set to 0.75, for the third word to 0.50, and for the fourth word, to 0.25.

A stepwise regression model fitted to the resulting simulated latencies suggested the same pattern of results, with a significant facilitatory effect only for n-gram frequency ($p = 0.044$). However, a generalized additive model (Wood, 2006) indicated the presence of a complex interaction of n-gram frequency by fourth word frequency, that we modeled with the help of a tensor product. The result is visualized in Figure 1, and Table 1 lists the statistics supporting the degrees of freedom invested in the tensor product. Across the full range of fourth word frequencies, we see a facilitatory effect of n-gram frequency. The effect of fourth word frequency, for a fixed n-gram frequency, is non-linear, and inverse U-shaped. The greatest simulated latencies are predicted for intermediate fourth word frequencies. If the weights of successive meanings indeed decrease with each additional word, then the prediction is that this interaction should also be present in the data of Arnon and Snider (2010). As they do not provide mean phrase decision latencies, we were not able to follow up on this prediction.

Model complexity

In order to evaluate the complexity of the NDR, we compare it to an (unimplemented) interactive activation model that includes representations for n-grams. There are two important aspects of model complexity, first, the complexity of the calculations involved, and second, the number of representations and connections required.

We first consider the complexity of the calculations. In the naive discriminative learning framework, the model's predictions are based on one forward pass of activation. Activations of meanings are obtained by summation over incoming active connections. An implicit assumption, shared with other computational models, is that there is a checking procedure that allows only those meanings to remain active that are supported by the input. Finally, a phrasal decision is based on the sum of activated meanings.

By contrast, an interactive activation model requires multiple cycles in which activation flows across inhibitory and excitatory connections at several layers. At each

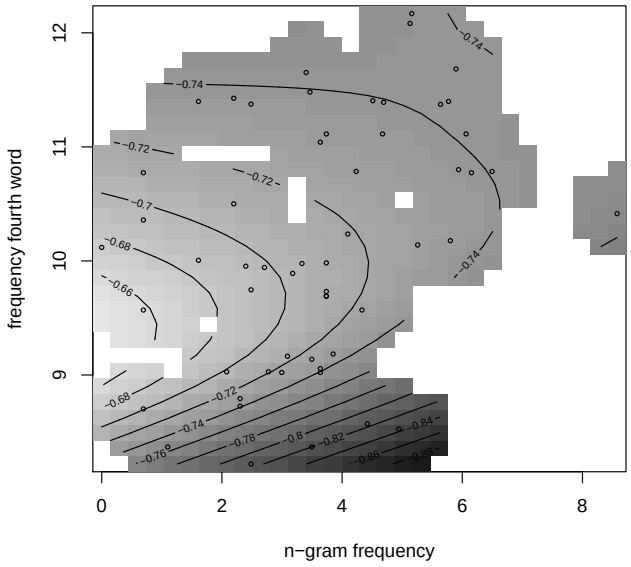

Figure 1: Regression surface estimated for the simulated latencies using a task-specific proportional weighting scheme of meanings. Lighter shades of gray indicate greater response latencies. Frequencies are shown on the log scale.

timestep, for each unit, the information coming in over its connections has to be evaluated. As for the NDR, a unit reaching a threshold activation level needs to be subjected to a checking procedure verifying that this unit is indeed compatible with the input. We maintain that the multiple cycles of interactive activation imply more complex computations than the single, and more local, forward pass of activation required for the NDR.

Next, consider the requirements of the models in terms of representations and connections. The present implementation of the naive discriminative reader requires 620 orthographic input units (letters and letter bigrams), 7494 meaning units, and $620 \times 7494 = 4{,}646{,}280$ connections from orthographic units to meaning units. The total number of units and connections is 4,654,394.

For an interactive activation model with n-gram representations, we derive the following estimates.

Given our lexicon, it turns out that there are 1,628,458 distinct n-grams ($1 \leq n \leq 4$). We assume that each of these n-grams does not spell out its component words, but provides pointers to its component words. For the present data set, 4,750,180 such pointers are required. Here, we ignore the possibility that each n-gram is also linked to its subordinate and superordinate n-grams. Finally, each distinct n-gram is associated with a frequency counter (or resting activation level), adding another 1,628,458 numeric representations.

Table 2: Model complexity evaluated in terms of counts of representations and connections, for an interactive activation model (IA) and the naive discriminative reader (NDR).

	IA	NDR
n-gram representations	1628458	0
n-gram-to-word links	4750180	0
n-gram frequency counters	1628458	0
letters	26	27
letter bigrams	0	593
meanings	7494	7494
word-to-meaning links	21146	0
word-to-letter links	168686	0
orthography-to-meaning links	0	4646280
total	8204448	4654394

We assume that words can be represented simply in terms of letters (26), without requiring letter bigrams. We also assume that, like the NDR, the interactive activation model is decompositional, and that hence the same 7494 meanings can be used to represent the meanings associated with an n-gram. In an approach in which syntactic n-grams receive separate representations, morphologically complex words should also have their own representations. Of the 1,628,458 distinct n-grams, 21,146 are distinct words (unigrams). We assume that each unigram provides links to its constituent letter representations. For the present set of 21,146 words, it turns out that 168,686 such connections from words to letters are required.

Table 2 summarizes the counts of model units (representations, links) in an interactive activation model and in the NDR. The total count for the interactive activation model is almost twice that for the NDR. We note here that the count for the interactive activation model is a lower bound if larger n-grams are also linked to lower-order n-grams with $1 < n < 4$.

An important difference that does not emerge from these counts is that the NDR is linear in the number of meanings: For every additional meaning, exactly 620 additional links from letter unigrams and bigrams to that meaning are required, in all, 621 model units. For the interactive activation model, each additional n-gram requires an additional representation, as well as additional links to its constituent words. Since there are hundreds of millions of n-grams, an interactive activation approach will require hundreds of millions of model units at the least, whereas 10 million such units is probably a generous upper bound for the NDR. We therefore conclude that naive discriminative learning provides a simpler and hence superior explanation of frequency effects above the (simplex) word level.

Discussion

We have shown that phrasal frequency effects in the lexical or phrasal decision task can arise as a straightforward consequence of naive discriminative learning. We note here that the explanatory potential of discriminative learning goes beyond lexical decision: for language acquisition, see Ramscar, Yarlett, Dye, Denny, and Thorpe (2010), and for the modeling of word naming latencies, see Hendrix and Baayen (2010).

Importantly, frequency effects in the NDR are inherently contextual in nature (Baayen, 2011b). In other words, frequency as pure repetition, such as might be modeled by a 'counter in the head' (such as a resting activation level for an n-gram representation) has no predictivity whatsover for the processing costs estimated by discriminative learning.

From the perspective of the NDR, interactive activation models ignore (or avoid) the problem of learning but pay the price of having to calculate the probabilities of a combinatory explosion of n-gram representations on-line, re-enacting for each n-gram token the learning process the adult state of which is represented in the associative weights of a naive discriminative learning network.

Naive discriminative learning should also be distinguished from subsymbolic connectionist modeling. In our model, the input and output layers contain symbolic, not subsymbolic units. Furthermore, instead of having to decide on network architecture (number of (sets of) hidden units, decay parameters, thresholds, etc.), the model has a fixed network layout, and in its simplest form (used here) there are no free parameters. The connection weights are obtained straightforwardly (and deterministically) from a corpus, by solving a system of equations.

The NDR can be viewed as a statistical classifier that is optimal in the least-squares sense, and that is grounded in well-established principles of animal and human learning. Baayen (2011a) shows that naive discriminative learning, used as a classifier for data on the dative alternation, actually outperforms a generalized linear mixed model, and performs as well as a support vector machine.

Given the present results, the existence of frequency effects for n-grams does not provide compelling evidence for the existence of separate representations for n-grams. Such representations may be required on independent grounds, as in the theory of Bod (2006). However, given the hundreds of millions of different n-grams even for small n ($n <= 7$), the huge costs of storage, retrieval of, and competition between putative millions of n-gram representations cannot be underestimated.

Instead of investing in n-gram representations, it may be fruitful to explore whether a hierarchy of naive discriminative learning networks can give rise to structured semantic representations, replacing the unstructured sets of meanings that our current implementation works with.

References

Arnon, I., & Snider, N. (2010). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Journal of Memory and Language*, *62*, 67–82.

Baayen, R. H. (2011a). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics*, *submitted*.

Baayen, R. H. (2011b). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *in press*.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual route model. *Journal of Memory and Language*, *36*, 94–117.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., & Marelli, M. (2010). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Under revision*.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*(3), 291–320.

Danks, D. (2003). Equilibria of the Rescorla-Wagner model. *Journal of Mathematical Psychology*, *47*(2), 109–121.

Hendrix, P., & Baayen, R. (2010). The Naive Discriminative Reader: a dual route model of reading aloud using naive discriminative learning. *Manuscript, University of Alberta*.

Milin, P., Filipović Durđević, D., & Prado Martín, F. Moscoso del. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, *60*(1), 50–64.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language. acquisition and communication* (pp. 151–173).

Wagner, A., & Rescorla, R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii* (pp. 64–99). Appleton-Century-Crofts.

Wood, S. N. (2006). *Generalized additive models*. New York: Chapman & Hall/CRC.